Recognition and Classification of Clinically Dysplastic Nevi from Photographs: A Study of Interobserver Variation

Patricia Hartge,[1] Elizabeth A. Holly, Allan Halpern, Richard Sagebiel, DuPont Guerry, David Elder, Wallace Clark, Louise Hanson, Christina Harrison, Robert Tarone, and Margaret A. Tucker

National Cancer Institute, NIH, Department of Health and Human Services, Bethesda, Maryland 20892 [P. H., R. T., M. A. T.], University of California at San Francisco Medical Center, San Francisco, California 94109 [E. A. H., R. S.], University of Pennsylvania, Philadelphia, Pennsylvania 19104 [D. G., D. E., W. C.], Westat, Inc., Rockville, Maryland 20850 [L. H., C. H.]

Abstract

The recognition of dysplastic nevi from photographs can aid in population surveys of nevi and in epidemiological studies of melanoma risk. The reproducibility of techniques for recognizing nevi as dysplastic or for scoring them according to the degree of dysplasia has not been measured. Using photographs of 300 nevi taken in the course of a case-control study of melanoma, we assessed the agreement among six clinicians in independently categorizing nevi as dysplastic and in grading the degree of dysplasia. On average, reviewers agreed with each other 77% of the time in classifying a nevus as dysplastic or normal. Pairwise agreement within one point on a six-point scale occurred 87% of the time on average. These results suggest that criteria for recognizing nevi as clinically dysplastic from photographs can be applied reproducibly.

Introduction

In 1978, distinctive melanocytic lesions were noted in 37 patients from six melanoma-prone families (1). Studies of these lesions [variously called dysplastic nevi (2), familial atypical mole and melanoma syndrome (3), or atypical moles (4)] have helped delineate their histological and clinical characteristics and their relationship to melanoma risk in people with and without a familial tendency. The current status of this research suggests a paradox; these clinically dysplastic nevi are strongly and consistently related to the risk of melanoma in epidemiological studies (5–9), yet controversy persists over their defining clinical and histological characteristics (4). When numerous severely atypical nevi occur, most observers recognize the phenomenon although they may disagree on the appropriate designation. Solitary atypical nevi and less severely atypical nevi arouse more controversy.

Various further studies are needed to resolve this contradiction, including studies of the reliability of histological (10, 11) and clinical diagnoses. We therefore studied the variability among six observers in the diagnosis of clinically dysplastic nevi from photographs taken during a case-control study of malignant melanoma.

Materials and Methods

This study of agreement among six experienced clinicians in the photographic evaluation of clinically dysplastic nevi was conducted as part of a case-control study of malignant melanoma conducted in clinics at the University of Pennsylvania and the University of California at San Francisco. Cases were patients newly diagnosed with malignant melanoma during a 2-year period that began in 1991. Controls were patients seen at other University of Pennsylvania and University of California clinics for conditions unrelated to the exposures under study.

During the case-control study, dermatology fellows specializing in melanoma (San Francisco) or nurse examiners (Philadelphia) saw all study subjects, examined their skin, and counted nevi on the entire skin surface except for on the perineum and scalp. The examiners selected the three most atypical nevi (even if these were deemed ordinary) of the subjects and photographed each one twice at 1:1 or 2:1 magnifications. A few subjects had fewer than three moles on the entire body, all of which were photographed. The entire back of each subject was photographed at a distance of approximately 6 feet. All photographs were developed in batches at a central lab at the National Cancer Institute and reviewed by the principal investigator (M. A. T.). A color control template was photographed as the first frame on each roll of film.

Subjects were chosen for this interobserver study at random. Photographs from cases and controls were interspersed. All photographs from each subject were shown sequentially. We projected onto two screens the pair of slides from one nevus and asked each reviewer to record his evaluation on a form that rated quality of the photograph, size of the nevus, and a summary evaluation of the nevus based on four characteristics (flatness, color, outline, and border).

The characteristics were recorded on 4-point scales: flatness (1 = all raised, 4 = all flat); color (1 = uniform, 4 = highly variegated); outline (1 = completely symmetric, 4 = highly asymmetric); and border (1 = completely distinct, 4 = completely indistinct). The reviewer also recorded a summary evaluation (not dysplastic, possible, mild, moderate, severe, morphologically dysplastic but less than 5 mm, melanoma, and can't tell). The criteria for classification of a nevus as dysplastic were: minimum size of 5 mm, a flat component, and at least two of the three features of variable color, asymmetric outline, or indistinct borders. Those that appeared morphologically dysplastic but smaller than 5 mm were classified as not dysplastic. The

Received 5/6/94; revised 10/18/94; accepted 10/19/94.

[1] To whom requests for reprints should be addressed at National Cancer Institute, 6130 Executive Blvd., EPN 443, Rockville, Maryland 20892.

Table 1 Consensus opinion of classification of individual nevi,[a] according to case-control status

Diameter	Dysplasia score	No. of nevi				%	
		From cases		From controls		Among ≥5 cm	
		Phila.[b]	S. F.	Phila.	S. F.	Case	Control
<5		14	28	43	66		
≥5	None	9	17	16	12	31	44
	Possible	4	9	8	11	15	30
	Mild	14	9	9	4	27	20
	Moderate	13	2	1	1	18	3
	Severe	1	2	1	0	4	1
	Melanoma	1	0	0	0	1	0
	Uncertain	1	3	1	0	5	1
All		57	70	79	94	100	100

[a] Excludes nevi with inadequate photographs.
[b] Phila., Philadelphia; S. F., San Francisco.

gradation of "mild, moderate, or severe" was not defined and was left to each reviewer.

The reviewers were not told the opinion recorded at the exam, whether the photographs were from a case or a control or the study center. The reviewers completed this form independently and without discussion, and the forms were collected before the slides of the next nevus were shown. After all three nevi were evaluated, the reviewers discussed each mole and reached a consensus on the classification of the person. Nevi for which neither photograph was adequate were excluded from analysis. The review sessions lasted about 30 h in total.

Reviewers evaluated photographs from 54 randomly sampled cases (26 from Philadelphia and 28 from San Francisco) and 68 random controls (32 from Philadelphia and 36 from San Francisco). In total, 300 photographed nevi were reviewed. Two of the reviewers were not available to review 23 of the nevi.

We analyzed differences in classification of nevi among reviewers using descriptive statistics, ANOVA and linear regression models. We estimated the intraclass correlation coefficient (12) as a measure of the agreement among the reviewers.

Results

About one-half of the nevi reviewed were smaller than 5 mm in diameter. As expected, the distribution of nevi varied between cases and controls and between Philadelphia and San Francisco (Table 1). On the basis of the consensus of the reviewers, 25 of the 79 (31%) nevi larger than 5 mm found on cases definitely were not dysplastic, 15% were possibly dysplastic, and 27% were only mildly dysplastic. In total, 18 nevi from cases (22%) and 3 from controls (4%) were deemed moderately or severely dysplastic (14%).

Some of the subjects selected for review had more than one dysplastic nevus. By consensus, 14 of the 68 (21%) control subjects were judged to have at least 1 dysplastic nevus and 25 of the 54 (46%) case subjects had at least 1 dysplastic nevus. Before the reviewers discussed all the photographs and reached a consensus they independently evaluated each nevus. As shown in Table 2, the fraction of all large nevi deemed not to be dysplastic varied from 35% (Reviewer A) to 47% (Reviewer C).

Table 2 Percentage of large nevi[a] classified in each score according to reviewer

Dysplasia score	Reviewer					
	A	B	C	D	E	F
None	35	36	47	38	43	44
Possible	3	17	10	17	24	19
Mild	34	28	33	30	23	26
Moderate	13	17	8	11	7	10
Severe	3	1	1	3	2	1
Melanoma	1	1	1	1	1	1

[a] Excludes nevi with inadequate photographs or size less than 5 mm in diameter.

Table 3 Percentage agreement between each pair of reviewers in characterizing large nevi[a] as dysplastic or not

Reviewer	A	B	C	D	E	F
A	100	81	77	87	73	77
B		100	82	78	72	75
C			100	79	67	76
D				100	72	80
E					100	77
F						100

[a] Excludes nevi with inadequate photographs or size less than 5 mm in diameter.

Table 4 Percentage agreement between each pair of reviewers in characterizing large nevi[a] within 1 point on a 6-point scale[b]

Reviewer	A	B	C	D	E	F
A	100	88	87	96	88	87
B		100	86	91	88	83
C			100	87	81	85
D				100	87	87
E					100	87
F						100

[a] Excludes nevi with inadequate photographs or size less than 5 mm in diameter.
[b] Scale of dysplasia scores: none, possible, mild, moderate, severe, and melanoma.

Table 5 Analysis of variance in the classification of color, flatness, outline, border, and summary dysplasia score

Dependent variable	Mean square, by source of variation[a]			Intraclass correlation coefficient[b]
	Reviewer	Nevus	Error	
Flatness	1.7	3.8	.47	.55
Color	0.9	1.2	.18	.49
Outline	3.5	2.5	.26	.58
Border	1.4	3.0	.55	.44
Summary	2.7	5.4	.49	.63

[a] From two-way analysis of variance.
[b] From one-way analysis of variance (12).

We compared the summary evaluation given by each possible pair of reviewers in two ways. First, we combined the categories of mild, moderate, or severe dysplastic nevi into one group and not dysplastic or possibly dysplastic into another group. According to this dichotomy, pairwise agreement ranged from 67–87% for large nevi with an average of 77% (Table 3). The rate of agreement by chance

Table 6 Average and standard error of scored characteristics[a] of large nevi,[b] for all reviewers combined, according to consensus classification

Dysplastic nevus	Flatness		Color		Outline		Border	
	Average	SE	Average	SE	Average	SE	Average	SE
No	1.4	.05	1.6	.03	1.6	.04	1.4	.05
Possible	2.3	.08	1.8	.04	2.0	.05	2.0	.06
Mild	2.7	.05	2.0	.03	2.3	.04	2.6	.05
Moderate	2.7	.08	2.5	.06	2.7	.07	2.5	.07
Severe	2.7	.20	2.4	.18	2.9	.21	2.5	.19

[a] All characteristics were scored from 1–4.
[b] Nevi included were ≥5 mm with adequate photographs. Melanoma and unknown were excluded.

between the members of all possible pairs would have been 51% on average.

We also compared the reviewers' scaled classification of each large nevus. Some reviewers consistently graded lesions higher than did other reviewers, but rarely did a reviewer disagree by more than one point on the scale. That is, the discrepancies were often "mild" *versus* "moderate" but seldom "mild" *versus* "severe." Table 4 shows the scaled data comparing each reviewer to each other reviewer. Agreement between reviewers within 1 point on the scale occurred between 81 and 96% of the time, 87% on average.

For each of the characteristics and the overall score for the presence of dysplasia, we analyzed how much the opinions varied among individual moles and among the six reviewers in a two-way ANOVA model (Table 5). Opinions varied less among reviewers than among nevi for flatness, color, border, and summary dysplasia score. Only for the classification of outline did the variation among reviewers exceed the variation among nevi. The variability was statistically significant among reviewers and among nevi for all characteristics. From the one-way ANOVA model for each dependent variable, we estimated the intraclass correlation coefficients, which ranged from .44 to .63 (Table 5). This indicates that the opinion of one reviewer about a particular nevus was substantially correlated with the opinions of the other reviewers.

To assess which characteristics contributed most to distinguishing dysplastic from normal nevi, we compared the evaluation of each reviewer for flatness, color, outline, and border to their summary classification of the nevus. (The melanomas and unknown are excluded.) As seen in Table 6, the range in average scores from normal nevi to severely dysplastic nevi was slightly greater for the extent of flatness and symmetry of outline than for distinctness of border. Color variegation discriminated large normal nevi from severely dysplastic nevi less than did the other features. These nevus characteristics naturally tended to covary, but simultaneous adjustment for all of the characteristics in a linear regression model showed statistically significant differences remaining for the extent of flatness, symmetry of outline, and irregularity of border, but not for variegation in color.

Discussion

These data show that multiple observers, who independently classify nevi as clinically normal or dysplastic from photographs, agree with each other most of the time, *e.g.*, the observers independently reached the same judgment of nevi 5 mm in diameter or larger as dysplastic or not about 77% of the time. Using a scale from definitely not dysplastic to melanoma, they agreed with each other within 1 point about 87% of the time. By comparison, an interobserver study of non-Hodgkin's lymphoma diagnosis found 60% agreement among expert pathologists in classification of non-Hodgkin's lymphoma (11). Two pathologists evaluating specimens from the vaginal cervix and grading them on a 6-point scale from no atypia to grade 3 cervical intraepithelial neoplasia agreed with each other within 1 point 77% of the time (13).

These data cannot disclose whether other observers would agree, but they show the feasibility of defining the appearance of dysplastic nevi in a fashion that can be reproducibly applied by different observers. The study also cannot show whether the assessments predict the histological appearance or the clinical behavior of the nevi. We plan to examine interobserver variation in histological classification of nevi from this study.

Only nevi at least 5 mm in diameter were deemed dysplastic. Dysplastic nevi were intermediate between normal nevi and melanomas in extent of flatness, variegation of color, symmetry of border, and distinctness of outline. No single characteristic served to distinguish dysplastic from normal nevi. The discussions of nevi after independent review also reflected an integration of the features. Furthermore, the reviewers often judged the nevus to be somewhere between clearly not dysplastic, on the one hand, and moderately or severely dysplastic, on the other. Some of this gradation, doubtless, reflected the loss of information in a photograph compared to a clinical exam; some gradation is present even when a clinical exam is possible.

We observed that color variegation mattered less than the other characteristics in discriminating among types of nevi. Inexact exposure settings, with slight over- or underexposure, reduce the apparent variation in color but minimally affect the symmetry of the outline or the distinctness of the border. Color may well provide better discrimination in clinical exams than it did in photographs. Similarly, lighting can be better controlled in a clinical exam, and palpation can help clarify surface texture, border elevations, and presence or absence of overall elevation. In these respects, clinical evaluation of nevi is superior to photographic evaluation.

On the other hand, photographic evaluation permits standardization of the review and allows as much time and discussion as needed. In epidemiological studies such as this, both clinical and photographic evaluation provide useful data to relate the type of nevi to risk. In clinical practice, photographs can also be useful for the same reasons and for detecting future changes in nevi.

References

1. Reimer, R. R., Clark, W. H., Greene, M. H., Ainsworth, A. M., and Fraumeni, J. F., Jr. Precursor lesions in familial melanoma, a new genetic preneoplastic syndrome. J. Am. Med. Assoc., *239:* 744–746, 1978.

2. Greene, M. H., Clark, W. H. Jr., Tucker, M. A., Elder, D. E., Tuthill, R., Hamilton, R., and LaRossa, D. Precursor nevi in cutaneous malignant melanoma: a proposed nomenclature. Lancet, *2:* 1024, 1980.

3. Lynch, H. T., Frichot, B. C., and Lynch, J. F. Familial atypical multiple mole-melanoma syndrome. J. Med. Genet., *15:* 352–356, 1978.

4. Diagnosis and Treatment of Early Melanoma. Bethesda, MD: Reprinted from NIH Consensus Development Conference Statement, Jan. 27–29, 1992, 10 (1).

5. Mackie, R. M., Freudenberger, T., and Aitchison, T. C. Personal Risk-Factor Chart for Cutaneous Melanoma. Lancet, *2:* 487–490, 1989.

6. Holly, E. A., Kelly, J. W., Shpall, S. N., and Chiu, S. H. Number of melanocytic nevi as a major risk factor for malignant melanoma. J. Am. Acad. Dermatol., *17:* 459–468, 1987.

7. Nordlund, J. J., Kirkwood, J., Forget, B. M., Scheibner, A., Albert, D. M., Lerner, E., and Milton, G. W. Demographic study of clinically atypical nevi in patients with melanoma and comparison subjects. Cancer Res., *45:* 1855–1861, 1985.

8. Halpern, A. C., Guerry, D., Elder, D. E., *et al.* Natural history of dysplastic nevi. J. Am. Acad. Dermatol., *29:* 51–57, 1993.

9. Schneider, J. S., Moore, D. H., and Sagebiel, R. W. Melanoma predicted by clinically atypical moles. Lancet, *339:* 1492, 1992.

10. Duncan, L. M., Berwick, M., Bruijn, J. A., Byers, H. R., Mihm, M. C., Rilke, F., Cacchelli, N., Fitzpatrick, T. B., and Sober, G. Histopathologic diagnosis of dysplastic nevi: concordance among pathologists convened by the WHO Melanoma Program. Hum. Pathol., *22:* 313–319, 1991.

11. Dick, F., VanLier, S., Banks, P., Frizzera, G., Witrak, G., Gibson, R., Everett, G., Schuman, L., Isacson, P., O'Conor, G., Cantor, K., Blattner, W., and Blair, A. Use of the working formulation for non-Hodgkin's lymphoma in epidemiologic studies: agreement between reported diagnoses and a panel of experienced pathologists. J. Natl. Cancer Inst., *78:* 1137–1144, 1987.

12. Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, *16:* 4, 1951.

13. Sherman, M. E., Schiffman, M. H., Erozan, Y. S., Wacholder, S., Kurman, R. J. A proposal for reporting abnormal cervical smears based on the reproducibility of cytopathologic diagnoses. Arch. Pathol. Lab. Med., *116:* 1155–1157, 1992.